Journal of Biomedical Informatics 46 (2013) 286–293

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Harmonization and semantic annotation of data dictionaries from the Pharmacogenomics Research Network: A case study

Qian Zhu *, Robert R. Freimuth, Zonghui Lian, Scott Bauer, Jyotishman Pathak, Cui Tao, Matthew J. Durski, Christopher G. Chute

*Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA*

ARTICLE INFO

*Article history:*
Received 24 July 2012
Accepted 17 November 2012
Available online 29 November 2012

*Keywords:*
Data harmonization
Semantic annotation
Pharmacogenomics

ABSTRACT

The Pharmacogenomics Research Network (PGRN) is a collaborative partnership of research groups funded by NIH to discover and understand how genome contributes to an individual's response to medication. Since traditional biomedical research studies and clinical trials are often conducted independently, common and standardized representations for data are seldom used. This leads to heterogeneity in data representation, which hinders data reuse, data integration and meta-analyses.

This study demonstrates harmonization and semantic annotation work for pharmacogenomics data dictionaries collected from PGRN research groups. A semi-automated system was developed to support the harmonization/annotation process, which includes four individual steps, (1) pre-processing PGRN variables; (2) decomposing and normalizing variable descriptions; (3) semantically annotating words and phrases using controlled terminologies; (4) grouping PGRN variables into categories based on the annotation results and semantic types, for total 1514 PGRN variables.

Our results demonstrate that there is a significant amount of variability in how pharmacogenomics data is represented and that additional standardization efforts are needed. This represents a critical first step toward identifying and creating data standards for pharmacogenomics studies.



## 1. Introduction

As biomedical research becomes more collaborative, challenges that arise when exchanging data among research groups becomes more pronounced. One of the primary, yet most fundamental, challenges in exchanging and integrating data is to ensure that data is both semantically (i.e., variable names and values share common meanings) and syntactically (i.e., the data shares a common format) interoperable. Incompatibilities often arise as a result of differences in the way research groups define and represent data. Overcoming these barriers usually requires one-to-one mappings and transformations between data sets. A more scalable approach is to define and use data standards, which ensure that all data collected using the standards for both the same semantic meaning and syntactic representation. Such standards, however, can be difficult to define in rapidly evolving fields of study where the types of data and/or the relationships between them change frequently. In those cases, standardization usually occurs after a sufficiently large corpus of data has been collected and research methods begin to converge. This manuscript describes results from the first step of just such a standardization process. Specifically, we describe the results of a case study of data dictionary standardization from members of the Pharmacogenomics Research Network (PGRN) [1].

PGRN is a collaborative partnership of research groups funded by the U.S. National Institutes of Health to discover and understand how genome contributes to an individual's response to medication. PGRN sites conduct very large scope of research fields, from cardiovascular–pulmonary diseases (including arrhythmias, hypertension, hypercholesterolemia, and asthma), cancers (including breast and gastrointestinal tumors and childhood leukemias), neuropsychiatric disorders (including depression and addiction), to classic determinants of drug blood levels (pathways of absorption, distribution, metabolism, elimination, and transport) [2]. There have been more than 1000 published fundamental and clinical research studies contributing significantly to the scientific base of knowledge in pharmacogenomics [3,4], a trend that is expected to continue. However, traditional biomedical research studies and clinical trials are being conducted independently, and common and standardized representations for data are seldom used. This leads to heterogeneity in the collected data and it hinders data reuse, integration and meta-analyses across multiple datasets.

## 2. Motivation

The variety of disease phenotypes are studied in the PGRN, as well as differences in clinical systems in use at each PGRN site, lead

* Corresponding author.
*E-mail address:* zhu.qian@mayo.edu (Q. Zhu).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.11.004

to data that is heterogeneous, non-standardized, and institution-specific. This not only hinders data aggregation among collaborating sites on a given study, but also complicates or prevents secondary use of the data, e.g., in meta-analyses.

To help overcome these issues, we performed a survey of PGRN data dictionaries, which are repositories of information about the data collected for a given study. Data dictionaries describe the variables used to capture data, including their meaning, origin, usage, relationships to other variables, and format. The goals of this survey were to: (1) identify overlapping and non-overlapping variables in the PGRN data dictionaries and (2) propose standards that establish a common semantic meaning and syntactic representation for the data.

For example, Table 1 lists several variables, along with their definitions and permissible values, from the data dictionaries of two PGRN sites. All three fields exhibit considerable variation as a result of both intra- and inter-site differences. As an example of intra-site inconsistency, Site 1 defines two different variables to capture information about ethnicity of a subject's maternal grandmother, which have different names, definitions, and permissible values. Interestingly, although the meaning of the permissible values is the same for the two variables, representation of the data is different, i.e. one variable uses integers while the other uses text. Inter-site differences between Sites 1 and 2 are also evident, as different names and permissible values are used to define the same concept. Furthermore, and perhaps most significantly, while the name and description of the variables defined by Site 1 indicate the data represents ethnicity, the values are an admixture of both ethnic and racial categories. This results in a discrepancy between variable name/description and a list of values, which will complicate interpretation and integration of the data.

Concepts of race and ethnicity are distinct and well-defined. In addition, the U.S. Office of Management and Budget (OMB) established standards for reporting race and ethnicity information that are already widely used [5] (Table 1). PGRN Site 2 conforms to the OMB standard but it employs a custom coding scheme and it lacks explicit definitions for the variables. This example illustrates how data consistency and comparability would be improved if both PGRN sites used the same definition and representation for common concepts. While this is only a simple example, it is common to find similar issues with other variables. In general, we have found that data heterogeneity tends to increase with the complexity of the data, the degree to which local coding systems are used, and the level of informality of the data dictionary. The harmonization effort described in this case study represents a critical first step toward identifying and creating data standards for pharmacogenomics studies.

## 3. Materials and methods

In this paper we demonstrate our approach to harmonize the data dictionaries of PGRN, which is a highly diverse research network. It emphasizes semantically annotating PGRN variables using the controlled terminologies, where possible, to avoid unnecessary proliferation of proposed standards in the biomedical research community. As shown in Fig. 1, we accomplished this task including multiple steps: (1) pre-processing PGRN variables; (2) decomposing and normalizing variable descriptions; (3) semantically annotating words and phrases using controlled terminologies; (4) grouping PGRN variables into categories based on annotation results and semantic types.

**Table 1**
Example of heterogeneity in data dictionaries: representation of race and ethnicity.

| Origin | Variable name | Variable description | Permissible values |
|---|---|---|---|
| PGRN Site 1 | Race_matern_gm | Ethnic background of your biological maternal grandmother | −8 = Not Applicable<br>−1 = Unknown<br>1 = Caucasian (White)<br>2 = African American<br>3 = Hispanic<br>4 = Asian<br>5 = Native American<br>6 = Other |
| PGRN Site 1 | Mat_gm_eth | Maternal grandmothers ethnicity | White<br>Black<br>Hispanic<br>Native american<br>Asian<br>Unknown<br>Other<br>Not applicable |
| PGRN Site 2 | Race | (none provided) | 1 = American Indian or Alaska Native<br>2 = Asian<br>3 = Black or African American<br>4 = Native Hawaiian or Pacific Islander<br>5 = White<br>6 = Unknown |
| PGRN Site 2 | Ethnicity | (none provided) | 1 = Hispanic or Latino<br>2 = Not Hispanic or Latino<br>3 = Unknown |
| OMB | Race | OMB race category (minimum designations) | American Indian or Alaska Native<br>Asian<br>Black or African American<br>Native Hawaiian or Pacific Islander<br>White |
| OMB | Ethnicity | OMB ethnicity category (minimum designations) | Hispanic or Latino<br>Not Hispanic or Latino |

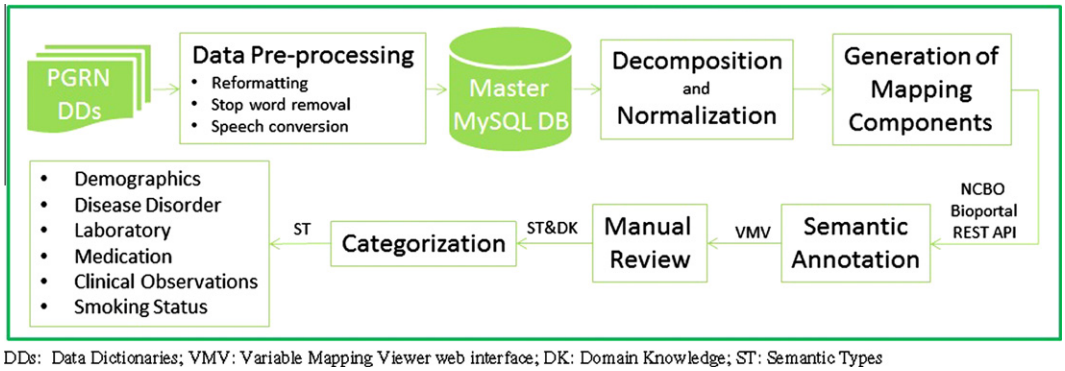


**Fig. 1.** Annotation pipeline used in this work.

### 3.1. Data pre-processing

Data dictionaries were collected from PGRN research sites. To accommodate differences in format, such as PDF, plain text, Microsoft Excel spreadsheets and html, and granularity of information provided for each variable, we pre-processed each dictionary by reformatting and filling in missing data, like missing variable descriptions or value sets. Discussions were held with the dictionary owner to obtain or clarify variable descriptions, value set contents, and define abbreviations. All variables were loaded into a MySQL database for harmonization. Each variable was assigned a unique identifier that was used throughout the entire harmonization process.

### 3.2. Decomposition and normalization

To provide consistent and comparable definitions for variables across research sites, terms from the controlled terminologies were used to capture semantic meaning of variable descriptions. As described below, NCBO Bioportal services were used to identify candidate terms. While the Bioportal service is designed to return both exact and partial matches, it is not designed to take long phrases, such as those typically found in data dictionaries, as input. Therefore, variable descriptions were decomposed and normalized for querying. For example, no annotation results were retrieved using whole phrase "Was the patient hospitalized for heart failure", even after stop words ("was", "the", "for") were removed. Therefore, we implemented an approach that is based on a lexical search algorithm. This approach first split each description into single words and short phrases, then removed stop words and normalized word form.

#### 3.2.1. Decomposition

Variable descriptions were first split into single words, which were then reassembled into phrases. The words and phrases, which we termed "mapping components" (MCs), were ultimately used as query terms for the Bioportal service. For instance, a description containing three words ("A B C") will produce seven MCs (A, B, C, AB, BC, AC, ABC). The length of each phrase was limited to a maximum of six single words.

#### 3.2.2. Stop word removal

Many words in variable descriptions are meaningless for semantic annotation. To improve results of the Bioportal queries, we removed all words that were contained in stop words list [6] and common English words list [7]. We also removed MCs including more than or equal to 50% stop words.

#### 3.2.3. Normalization

The level of formalism in data dictionaries varies greatly. To remove the colloquialism in variable definitions, speech conversion and tense normalization were implemented based on Unified Medical Language System (UMLS) Specialist Lexicon [8]. This process converted verb tense to a common base form, plural nouns to singular form, and possessive nouns to base forms using LRAGR lexicon. In addition, verbs, adjectives, and adverbs were converted to nouns using LRNOM lexicon.

Table 2 shows an example of a variable description that was decomposed and normalized. In this example, "was", "the", "for" were removed as stop words, and "was the", "the patient", "hospitalization for", "for heart", and "was the patient", etc. were removed due to the percentage of stop words meeting or exceeding 50%. In addition, "hospitalized" was converted to "hospitalize" by LRAGR, and then converted to "hospitalization" by LRNOM.

### 3.3. Semantic annotation and categorization

To complete semantic annotation process, MCs generated from the previous step were used to query controlled terminologies, results were reviewed manually, and UMLS semantic types (ST) [9] for the selected terms were used to group variables into different categories.

#### 3.3.1. Annotation with controlled terminologies

Based on types of data collected in pharmacogenomics domain, SNOMED-CT [10], NDF-RT [11], NCI Thesaurus [12], RxNorm [13] and LONIC [14] were selected as source terminologies for semantic annotation. NCBO BioPortal [15] provides access to many biomedical ontologies, including those selected for this study. An annotation pipeline was developed to utilize BioPortal Web services [16],

**Table 2**
Example of variable description decomposition and normalization.

| Original variable | Resulting Mapping Components (MCs) | |
|---|---|---|
| Was the patient hospitalized for heart failure | *Single words* | Patient, hospitalization, heart, failure |
| | *Phrases* | Patient hospitalization, heart failure, patient hospitalization for, hospitalization for heart, for heart failure, patient hospitalization for heart, hospitalization for heart failure, patient hospitalization for heart failure |

which provide programmatic access to terminology content. This annotation pipeline used for MCs obtained above to query five ontologies selected for this study. Query results were returned in XML format, which were loaded into a database for manual review.

### 3.3.2. Annotation review

Annotation results were manually reviewed to ensure that semantic meaning of each corresponding variable description was captured. To facilitate such review process, a simple web application was developed that allowed curators select the best term(s) for annotation (Fig. 2). The web application presented all of the terms that were returned for a given variable, using the variable's MCs as query terms. Curators reviewed each variable description and selected term(s) that were thought to best represent semantic meaning of such variable, as indicated by the check box in the "Accepted Mapping" column in Fig. 2.

Following term selection, curators determined how completely selected terms captured semantic meaning of the variable. Each variable was given a status of "complete mapping", "partial mapping" or "no mapping". Variables with status as "complete mapping" were used in the next step, variable categorization directly, while those variables that were not sufficiently represented by the query results were flagged for further study, e.g., additional clarification of the semantic meaning with the owner of the data dictionary or manual annotation.

### 3.3.3. Categorization

To facilitate harmonization process, variables were categorized into common domains, such as demographics, medications, and laboratory results. This was accomplished by taking advantage of mappings that exist between terminologies that were used for semantic annotation and UMLS semantic types (ST). UMLS ST are organized in a hierarchical tree. As shown in Fig. 3, "Disease or Syndrome" is a child node of "Pathologic Function", and "Disease or Syndrome" is a parent node of "Mental or Behavioral Dysfunction". ST hierarchical tree also allowed us to uniformly represent annotations at different levels of granularity, which is likely since different terminologies were utilized for annotation. For example, "Atrial Fibrillation" is a "Disease or Syndrome" in NCI Thesaurus but it is a "Pathologic Function" in SNOMED-CT and NDF-RT. ST hierarchy provides a means to identify a common category ("Pathologic Function") for "Atrial Fibrillation" across all three terminologies.

Several domains were chosen as variable categories, which were mapped to ST categories (Table 3), based on the types of variables that were present in data dictionaries used for this study. ST of a primary concept, which was determined by the manual review of semantic annotations was then used to categorize the variable into one of the domains. Such as, "past angina", the primary concept "angina" with disorder ST is used for categorization, and "past" is as temporal qualifier for "angina".

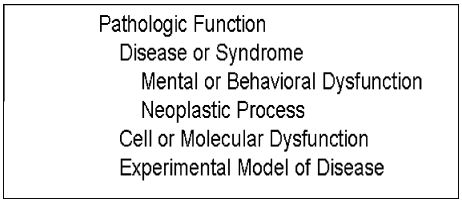


**Fig. 3.** Subset of UMLS semantic types hierarchical tree.

**Table 3**
Grouping UMLS semantic types into proposed domains.

| Domains | Relevant semantic types |
|---|---|
| Demographic | Organism attribute; organism function |
| Medication | Pharmacologic substance; clinical drug; organic chemical |
| Laboratory | Laboratory or test result; laboratory procedure |
| Disorder | Disease or syndrome; mental or behavioral dysfunction; pathologic function |
| Smoking status | Environmental effect of humans |
| Clinical observation | Clinical attribute |

## 4. Results

### 4.1. PGRN data dictionaries

A total of 1514 variables were collected from four PGRN sites. Following manual review, a number of variables were found to be highly specific and therefore less likely to be reused, or repeatedly used across dictionaries from same site. As shown in Table 4, 84 variables were classified as site-specific, many of which represented processing state or internal flags, e.g., "uploaded to database", "Field for Skip logic". A total of 65 variables were found to differ by only a time-based qualifier, e.g., blood pressure at visit 1, 2, or 3, and 514 variables were repeated across dictionaries. The latter category included instances of variables that were repeated to create a list, e.g., Drug 1 name, Drug 2 name, etc., and those that were identical copies in different dictionaries, thereby representing instances of variable reuse.

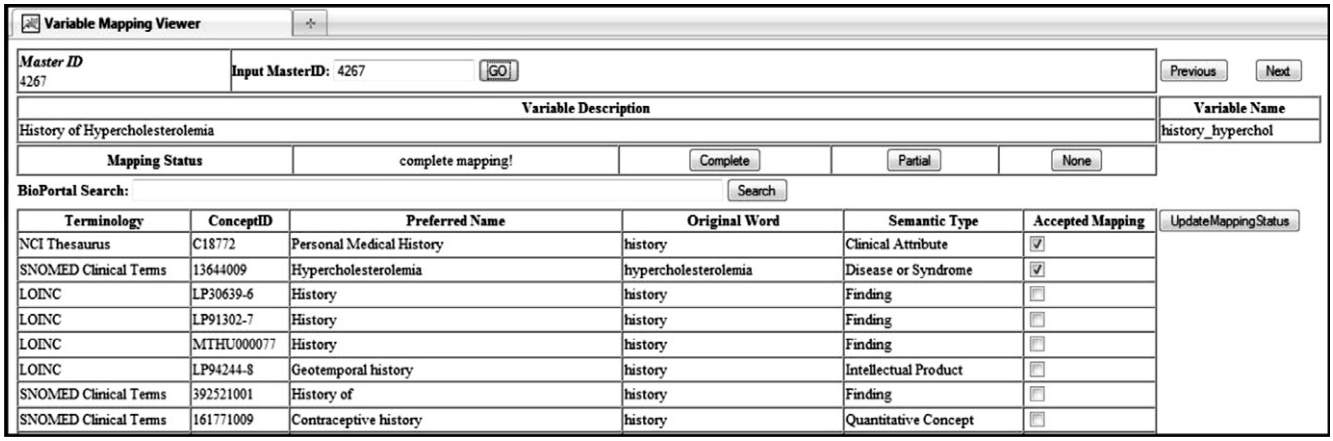


**Fig. 2.** Snapshot of "Variable Mapping Viewer" web interface.

**Table 4**
Number of special variables collected from PGRN sites.

| Type of variable | Descriptions | PGRN GROUP 1 | PGRN GROUP 2 | PGRN GROUP 3 | PGRN GROUP 4 | Total |
|---|---|---|---|---|---|---|
| Site-specific | Variables designed for internal use with site specific flags | 74 | 8 | 1 | 1 | 84 |
| Differ only by time qualifier | Variables designed for recording different results retrieved for one particular event (diagnosis, laboratory test, etc) at different time points | 5 | 1 | 59 | 0 | 65 |
| Repeated | Variables with same semantic meanings | 451 | 49 | 14 | 0 | 514 |
| Unique | Variables with different semantic meaning | 317 | 409 | 107 | 18 | 851 |
| Total | | 847 | 467 | 181 | 19 | 1514 |

**Table 5**
Decomposing and normalizing results.

| | PGRN GROUP 1 | PGRN GROUP 2 | PGRN GROUP 3 | PGRN GROUP 4 | Total |
|---|---|---|---|---|---|
| Total number of MCs | 7389 | 3827 | 1857 | 54 | 13,127 |
| Total number of MCs removed by stop words scanning | 1203 | 2247 | 520 | 0 | 3970 |
| Total number of MCs converted by specialist lexicon | 417 | 348 | 102 | 1 | 868 |

### 4.2. Decomposing and normalizing

Since variable names tend to be highly abbreviated and rarely capture the full semantics of the data that they represent, variable descriptions were chosen as a source for semantic annotation. To accomplish this, variable descriptions were decomposed into single words and short phrases, normalized, and then used as query terms to search controlled terminologies.

A total of 16,914 MCs were generated for 1514 variables used in this study (Table 5). As described above, stop words and phrases that contained at least 50% stop words were removed prior to executing the query. This step reduced the number of MCs by 3970. In addition, two Specialist Lexicons, LRAGR and LRNORM, were used for speech and tense conversion; consequently 868 MCs were converted to base forms.

### 4.3. Semantic annotation

#### 4.3.1. Annotated by controlled terminologies

MCs generated from above steps were annotated by controlled terminologies described in Section 3.3. Invoking NCBO Bioportal RESTful API, annotation results were generated and rendered in the XML format shown in Table 6. All results were reviewed manually to determine the most matched appropriate terms.

#### 4.3.2. Annotation review

Annotation results were reviewed by using the web application shown in Fig. 2. For example, three MCs: ''History'', ''Hypercholesterolemia'', and ''History of Hypercholesterolemia'' were generated for a variable description, ''History of Hypercholesterolemia''. Each of these MCs was used as a query term for searching the five aforementioned terminologies, results of which were reviewed by a curator. Term selection was based both on term definition as well as ST of the term. In this example, ''Hypercholesterolemia'' is a disease, so candidate terms that had a non-disease ST were excluded from consideration, and ''personal medical history'' with ''Clinical Attribute'' as ST was selected as the best term to represent the concept of ''history''. Terminology preference for specific domains was also considered as a determine factor when a given concept had mappings to multiple terminologies. Specifically, SNOMED-CT was preferred for representing concepts related to disease, RxNorm was preferred for representing concepts related to medications, and LOINC was preferred for representing concepts related to laboratory tests. Finally, the example shown in Fig. 2, ''History of Hypercholesterolemia'' was marked as a ''complete mapping'', since the semantic meaning of the variable was completely captured using the selected terms.

Two observations became evident during the annotation step. First, variables in this study were, in general, highly pre-coordinated and therefore they required several concepts to capture their semantic meaning. For example, it is common to record a subject's race in pharmacogenomics studies, since allele frequencies can vary widely among different racial groups. Furthermore, in family studies, it is common to record not only a primary subject's race, but also a race of family members. The data dictionaries used for this study included several variables that captured the race of different individuals, each of which was semantically identical at both level of the variable description (''race category'') and its permissible values, e.g., ''American Indian or Alaska Native'', ''Asian'', etc.; see Table 1, but which differed from each other due to the term that represented relationship of individual in question to that of the primary subject. In these cases, it is preferable to use a generic variables to represent the primary concept, and its set of permissible values, or value domain, then add a qualifier to capture the distinguishing factor. While this may be difficult to achieve on a case report form or family history questionnaire, it is relevant to the data models that are used to represent the information.

Secondly, many variables were captured as derived values, e.g., the age of the subject at diagnosis, the age of the subject at hospitalization, etc, rather than as primary data, e.g., birth date, date of

**Table 6**
Semantic annotation results.

| | PGRN GROUP 1 | PGRN GROUP 2 | PGRN GROUP 3 | PGRN GROUP 4 | Total |
|---|---|---|---|---|---|
| Total number of MCs | 7389 | 3827 | 1857 | 54 | 13,127 |
| Total number of mappings from five terminologies | 48,509 | 20,652 | 9683 | 673 | 79,517 |
| *Total number of mappings* | | | | | |
| LOINC | 12,308 | 4852 | 2409 | 158 | 19,727 |
| NCI Thesaurus | 10,719 | 4801 | 2328 | 152 | 18,000 |
| NDF-RT | 6244 | 2813 | 1315 | 106 | 10,478 |
| RxNORM | 6758 | 2838 | 1083 | 105 | 10,784 |
| SNOMED-CT | 12,480 | 5348 | 2728 | 152 | 20,708 |

**Table 7**
Categorization results and examples for 797 variables from four PGRN groups.

| Categories | # Variables | Examples | | | | | |
|---|---|---|---|---|---|---|---|
| | | Variables | MC | Preferred name | Concept code | Terminologies | ST |
| Medication | 170 | Drug Strength | Drug | Substance | C459 | NCI Thesaurus | Pharmacologic substance |
| | | | Strength | Pharmaceutical Strength | C53294 | NCI Thesaurus | Qualitative concept |
| | | Currently taking aspirin | Aspirin | Aspirin | 1191 | RxNorm | Organic chemical |
| | | | Currently | Current | 15240007 | SNOMED CT | Temporal concept |
| Disease disorder | 146 | Lone atrial fibrillation | Lone atrial fibrillation | Lone atrial fibrillation | 233910005 | SNOMED CT | Disease or syndrome |
| | | History of Myocardial Infarction | Myocardial Infarction | Myocardial Infarction | 22298006 | SNOMED CT | Disease or syndrome |
| | | | History | Personal Medical History | C18772 | NCI Thesaurus | Clinical attribute |
| Clinical observation | 71 | Clinic diastolic blood pressure | Clinic | Clinic | C51282 | NCI Thesaurus | Health care related organization |
| | | | Diastolic blood pressure | Diastolic blood pressure | 271650006 | SNOMED CT | Clinical attribute |
| Laboratory | 69 | Electrophysiology study | Electrophysiology | Electrophysiology | LP6252-3 | LOINC | Laboratory procedure |
| Smoking status | 65 | What age quit smoking | Age | Age | LP28815-6 | LOINC | Organism attribute |
| | | | Smoking | Tobacco Smoking | C17934 | NCI Thesaurus | Individual behavior |
| | | | Stop | Stop | C65125 | NCI Thesaurus | Activity |
| Demographics | 62 | Age | Age | Age | LP28815-6 | LOINC | Organism attribute |
| | | Gender | Gender | Gender | LP61312-2 | LOINC | Organism attribute |
| Other categories | 214 | DNA Sample Number | DNA | DNA | LP32416-7 | LOINC | Nucleic acid, nucleoside, or nucleotide |
| | | | Sample | Specimen | C19157 | NCI Thesaurus | Physical object |
| | | | Number | Number | C25337 | NCI Thesaurus | Quantitative concept |

diagnosis, and date of hospitalization. While it is convenient to capture derived values that are relevant for a particular study, it is more difficult to utilize data set for secondary purposes. Capturing data as primary values simplifies data integration and reuse.

### 4.4. Categorization

Only variables with "complete mapping" label were moved into this categorization step. We used the selected annotation results with ST information and relied on human domain knowledge to categorize variables into categories. The categories are shown in Table 7. Note that the variables included in Table 7 were calculated based on the 797 "unique" variables only (see Table 4).

It is not surprising that pharmacogenomics data sets contain a relatively large number of variables that represent medications, diseases, clinical observations, laboratory values, and demographics. However, it should be noted that many laboratory-based variables, such as "gamma-glutamyl hydrolase activity in diagnostic bone marrows" and "R enantiomer of the primary metabolite Desmethyl Citalopram (ng/mL)"could not be fully annotated and therefore categorized since there was no suitable term, e.g., LOINC code to represent. This may be due to the fact that some laboratory tests that are used in pharmacogenomics studies are conducted in experimental, rather than clinical, labs. As pharmacogenomics data is integrated into clinical practice, it may be necessary to extend terminologies to represent new laboratory tests.

It was also striking that none of the pharmacogenomics data dictionaries used in this study contained variables that represented genomic data. Obviously, the research sites that provided the dictionaries generate and store genomic data. The absence of these elements in their data dictionaries may be a reflection of the relative immaturity of the application of pharmacogenomics data in a clinical setting and a tendency to consider the genomic data experimental. The lack of standards to represent pharmacogenomics data may also be a factor. Clearly, this is an area for future work.

### 4.5. Evaluation

Domain experts inside Mayo Clinic were invited to review our semantic annotation work, including the annotation selections and categorization outcomes. Based on their evaluation results, we performed two further evaluations to determine overall performance of our harmonization infrastructure. Valuable evaluations by PGRN sites have not been done, but will take place in the coming months.

#### 4.5.1. Semantic annotation

In this evaluation step, we considered annotation results only for the "unique" variables without duplicated and repeated ones.

**Table 8**
Semantic annotation results.

| PGRN Groups | # Variables | # "complete mapping" | # "partial mapping" | # "no mapping" |
|---|---|---|---|---|
| PGRN GROUP 1 | 317 | 295 | 11 | 11 |
| PGRN GROUP 2 | 409 | 387 | 17 | 5 |
| PGRN GROUP 3 | 107 | 97 | 4 | 6 |
| PGRN GROUP 4 | 18 | 18 | 0 | 0 |
| Total # | 851 | 797 (93.6%) | 32 (3.8%) | 22 (2.6%) |

**Table 9**
Categorization results with semantic types.

| | Demographics | Medication | Laboratory | Disease disorder | Clinical observation | Smoking status |
|---|---|---|---|---|---|---|
| PGRN GROUP 1 | 9 (69.2%) | 85 (100%) | 31 (72.1%) | 28 (87.5%) | 47 (83.9%) | 5 (83.3%) |
| PGRN GROUP 2 | 23 (67.6%) | 45 (84.9%) | 11 (57.9%) | 85 (96.6%) | 4 (75%) | 55 (93.2%) |
| PGRN GROUP 3 | 6 (54.5%) | 24 (100%) | 4 (80%) | 22 (100%) | 8 (100%) | 0 (100%) |
| PGRN GROUP 4 | 3 (75%) | 8 (100%) | 2 (100%) | 4 (100%) | 0 (100%) | 0 (100%) |
| Total | 41 (66%) | 162 (95.3%) | 48 (69.6%) | 139 (95.2%) | 59 (83.1%) | 60 (92.3%) |

Table 8 shows that 93.6% PGRN variables in this study can be fully captured by the annotation results selected by curators. The number of complete annotations can increase by performing additional modifications for the variables with partial/no mapping.

*4.5.2. Categorization with semantic types*

A total of 583 variables were grouped into six categories based on semantic types and domain knowledge. The matched results displayed as numbers along with percentages are shown in Table 9. From Table 9 and 509 variables (87.3%) have been successfully grouped into appropriate categories by ST, and 74 (12.7%) variables were not placed in any relevant categories by ST. Main reason of the 12.7% failure is a primary word missing in such variables, resulted in no corresponding ST assigned for these variables, such as "dose", "Dosing frequency", etc., which are missing "drug" as primary word. For such cases, we manually moved them into correct groups.

## 5. Limitation and future work

Variables from PGRN sites were not distinguished with value sets completely, that is to say, some variables were value sets. For example, we had "subject race" and "American Indian or Alaskan Native" as individual variables, and the second one should be the value set of the first one "subject race". In this work, we did not differentiate these variables and process them separately, but in future work we will extract value set from the mixed data sets and combine permissible values provided by PGRN sites separately, and then standardize and load them into LexEVS [17] for future browsing and querying.

We aggregated and processed data from four PGRN groups, and generated six common categories in this work. However, the workflow reported in this paper will be used to handle datasets from more PGRN sites; and undoubtedly, more categories will be generalized on the basis of particular research focuses from these sites. Meanwhile, site-specific variables will be taken into account in future work.

Due to a huge portion of PGRN clinical data received currently, in this study, we were focusing on clinical data processes, which are relevant to laboratory test, medication, and disease. Meanwhile, we did collect some genomics sample data from particular PGRN groups, and we expect more genomics data descriptors will be able to be placed into our PGRN data repository in near future. Then we will collaborate with a joint Genomics Work Group, established by HL7 [18] and CDISC [19] to address problems associated with genomics data harmonization and generate PGRN specific genomics data standards.

To fill a gap between pharmacogenomics data standardization, linkage to Electronic Medical Record (EMR) and clinical research standards, further mappings with standardized clinical data models for each category will be taken into account. We propose to map PGRN variables from each category to Clinical Element Model [20], CDISC [19], Case Report Forms from caDSR [21], and PhenX [22]. This future work will not only make PGRN variables representable in a more standardized way, but also provide flexibility of bridging and expanding PGRN specific variables to the clinical data models.

## 6. Conclusion

Data and metadata standards help to mitigate problems that arise from semantic and syntactic differences between research groups. These differences are major barriers that hinder effective communication among scientists and that slow the pace of advancement and discovery. It is often difficult for those in rapidly advancing fields of study to converge on a set of standards before a significant volume of data is generated. This can result in the generation of large data sets that are difficult to interpret, merge together, and use in downstream analyses that were not part of the original study design. This work describes initial effort to harmonize data dictionaries from pharmacogenomics research sites. Our results demonstrate that there is a significant amount of variability in how data is represented among PGRN sites and that a larger standardization effort is needed.

## Acknowledgment

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomic Research Network).

## References

[1] PGRN. <http://pgrn.org/display/pgrnwebsite/PGRN+Home>; 2012 [accessed July 2012].
[2] Long RM, Berg JM. What to expect from the pharmacogenomics research network. Clin Pharmacol Ther 2011;89:339–41.
[3] O'Donnell Peter H, Ratain Mark J. Germline pharmacogenomics in oncology: decoding the patient for targeting therapy. Mol Oncol 2012.
[4] Abo Ryan, Hebbring Scott, Ji Yuan, Zhu Hongjie, Zeng Zhao-Bang, Batzler Anthony, Jenkins Gregory D, Biernacka Joanna, Snyder Karen, Drews Maureen, Fiehn Oliver, Fridley Brooke, Schaid Daniel, Kamatani Naoyuki, Nakamura Yusuke, Kubo Michiaki, Mushiroda Taisei, Kaddurah-Daouk Rima, Mrazek David A, Weinshilboum Richard M. Merging pharmacometabolomics with pharmacogenomics using '1000 Genomes' single-nucleotide polymorphism imputation: selective serotonin reuptake inhibitor response pharmacogenomics. Pharmacogenet Genom; 2012.
[5] The OMB defines race and ethnicity as two categories, and prefers that this data be collected separately. The OMB does allow a combined format, however, that combines the permissible values from each category. <http://www.census.gov/population/www/socdemo/race/Ombdir15.html>; 2012 [accessed July 2012].
[6] Stop words. <http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/>; 2012 [accessed July 2012].
[7] Common english words. <http://www.textfixer.com/resources/common-english-words.php>; 2012 [accessed July 2012].
[8] UMLS Specialist Lexicon. <http://www.nlm.nih.gov/pubs/factsheets/umlslex.html>; 2012 [accessed July 2012].
[9] UMLS Semantic Types. <http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html>; 2012 [accessed July 2012].
[10] SNOMED CT. <http://www.ihtsdo.org/snomed-ct/>; 2012 [accessed July 2012].
[11] Brown SH, Elkin PL, Rosenbloom ST, Husser C, Bauer BA, Lincoln MJ, et al. VA national drug file reference terminology: a cross-institutional content coverage study. Medinfo 2004:477–81.
[12] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Infor 2007;40(1):30–43.
[13] Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof 2005;7(5):17–23.

[14] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003;49:624–33.
[15] Rubin DL, Moreira DA, Kanjamala PP, et al. BioPortal: a web portal to biomedical ontologies. Assoc Adv Artif Int 2007.
[16] NCBO REST API. <http://www.bioontology.org/wiki/index.php/NCBO_REST_services>; 2012 [accessed July 2012].
[17] LexEVS. <https://wiki.nci.nih.gov/display/LexEVS/LexEVS>; 2012 [accessed July 2012].
[18] HL7. <http://www.hl7.org/>; 2012 [accessed July 2012].
[19] CDISC. <http://www.cdisc.org>; 2012 [accessed July 2012].
[20] Clinical Element Model. <http://intermountainhealthcare.org/CEM/>; 2012 [accessed July 2012].
[21] caDSR CRF. <https://formbuilder.nci.nih.gov/FormBuilder/>; 2012 [accessed July 2012].
[22] Stover PJ, Harlan WR, Hammond JA, et al. PhenX: a toolkit for interdisciplinary genetics research. Curr Opin Lipidol 2010;21:136–40.